Optimization over time-varying directed graphs with row and column-stochastic matrices

Fakhteh Saadatniaki, *Student Member, IEEE*, Ran Xin, *Student Member, IEEE*,
and Usman A. Khan†, *Senior Member, IEEE*

***Abstract*—In this paper, we provide a distributed optimization algorithm, termed as TV-$\mathcal{AB}$, that minimizes a sum of convex functions over time-varying, random directed graphs. Contrary to the existing work, the algorithm we propose does not require eigenvector estimation to estimate the (non-1) Perron eigenvector of a stochastic matrix. Instead, the proposed approach relies on a novel information mixing approach that exploits both row- and column-stochastic weights to achieve agreement towards the optimal solution when the underlying graph is directed. We show that TV-$\mathcal{AB}$ converges linearly to the optimal solution when the global objective is smooth and strongly-convex, and the underlying time-varying graphs exhibit bounded connectivity, i.e., a union of every C consecutive graphs is strongly-connected. We derive the convergence results based on the stability analysis of a linear system of inequalities along with a matrix perturbation argument. Simulations confirm the findings in this paper.**

***Index Terms*—Optimization, distributed algorithms, first-order methods, time-varying graphs, directed graphs.**

I. Introduction

With the advent of 5G, promising higher bandwidth and faster data rates, emerging technologies like the Internet of Things, self-driving cars, and smart devices are coming to the forefront. In these applications, it is of paramount interest to learn hidden parameters from the data collected by the individual units [1]. For example, self-driving cars rely heavily on computer vision in order to correctly identify pedestrians, highway lanes, or traffic signs in all light and weather conditions. Problems such as these can be framed as classification, regression, or risk minimization, at the heart of which is a simple sum of cost optimization. The sheer size of data and privacy concerns limit data sharing and thus solutions to the underlying optimization problems must be developed that are local and distributed [2]–[4]. However, departing from the traditional approaches, what is needed now are algorithms that are applicable to mobile and autonomous agents resulting in a time-varying and non-deterministic information exchange.

Recently, there has been a large body of work on distributed optimization, where the goal is to minimize a sum of costs:

$$\min_{\mathbf{x}} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}),$$

such that each objective function, $f_i : \mathbb{R}^p \to \mathbb{R}$, is private to agent i. In order to solve this problem, the agents exchange information with nearby nodes over a sparse communication graph. Such problems arise for example in sensor networks [5], large-scale machine learning [6], [7], distributed estimation [8], and localization [9]. When the graphs are static and undirected, early work on first-order methods include [10]–[12] with the convergence rate of $O(\frac{\ln k}{\sqrt{k}})$ for arbitrary convex functions and $O(\frac{\ln k}{k})$ for strongly-convex functions, where k denotes the number of iterations. The sublinear convergence is due to the use of diminishing step-sizes. The rate improves to linear with a constant step-size but at the expense of a sub-optimal solution [13]. Methods based on Lagrangian dual [14], [15] converge faster but suffer from a high computational burden as they require solving a subproblem at each iteration.

Optimization over directed graphs is developed in [16]–[18], where push-sum [19], [20] is used to achieve consensus among the agents. The convergence rate, with diminishing step-sizes, is $O(\frac{\ln k}{\sqrt{k}})$ for arbitrary convex functions and $O(\frac{\ln k}{k})$ for strongly-convex functions. In contrast, Refs. [21], [22] use an alternate approach called surplus consensus [23] to achieve consensus but with the same convergence rates as [16]–[18]. The main reason for slow convergence is that a local gradient is used at each agent, which requires diminishing step-sizes to ensure optimality. To overcome this challenge, Refs. [24]–[26] replace the local gradient with an estimate of the global gradient, with the help of dynamic consensus [27] over undirected graphs, and show linear convergence to the optimal solution. This gradient estimation approach was combined with push-sum (type) methods in [26], [28]–[30] to achieve linear convergence to the optimal solution over directed graphs. Related work along these lines also include [31] over undirected graphs and [32] over directed graphs that are related to [26], [28]. An alternate approach that builds on [21], [22] and does not use push-sum has been recently developed in [33], where a row- and a column-stochastic matrix are used simultaneously to achieve linear convergence to the optimal over directed graphs. Accelerated methods can be found in [34]–[36], while non-convex problems are considered in [37].

In this paper, our focus is on time-varying directed graphs. Early work on time-varying communication among the agents can be found in [38], [39], where the exchange is undirected and in [40], where the exchange is directed. These methods are built on local gradients at each agent and are thus sublinear. Recent work includes [41], where the authors establish a geometrically converging distributed optimization algorithm over directed graphs under uncoordinated yet bounded step-sizes, and [42], where agents communicate over graphs subject to random link failures. Work on random graphs can be found

†The authors are with the Department of Electrical and Computer Engineering at Tufts University, Medford, MA 02155; {fakhteh.saadatniaki@,ran.xin@,khan@ece.}tufts.edu.
This work is partially supported by an NSF Career award: CCF # 1350264.

in [43], where the problem of constrained convex optimization is investigated for non-differentiable costs under Markovian communication model. For random networks modeled by a sequence of independent, identically distributed (IID) random matrices drawn from the set of symmetric, stochastic matrices with positive diagonals, [44] proposes two accelerated distributed Nesterov-like gradient methods featuring resiliency to link failures, reduced computational load, and improved convergence rates compared to other gradient methods. Ref. [45] establishes a convergence rate of $O(\frac{1}{k})$ for distributed stochastic gradient methods over temporally IID random undirected graphs for strongly-convex costs when local gradients are subject to noise that is IID in time and has a finite second moment. Furthermore, asynchronous multi-agent optimization is considered in [46], where the authors adapt the curvature estimation technique of the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton optimization method [47]–[49] for use in asynchronous distributed settings over undirected graphs [50]. An asynchronous implementation of subgradient-push [18] algorithm is also recently developed in [51].

In this paper, we use gradient estimation as was used in [24]–[26], [28] and extend the $\mathcal{AB}$ algorithm introduced in [33] to time-varying, random directed graphs. Of relevance in this context is Ref. [26], which uses gradient estimation and push-sum consensus [19], [20] to implement distributed optimization over time-varying graphs. However, the push-sum based methods [26], [28]–[30] involve estimating the (non-$\mathbf{1}$) Perron eigenvector of a stochastic matrix and an appropriate scaling with the estimated eigenvector components. The eigenvector estimation adds conservatism to the overall algorithm as its convergence may dominate the overall rate. In contrast, the $\mathcal{AB}$ algorithm [33] does not require such eigenvector estimation and employs both row- and column-stochastic weights in a novel way. The time-varying algorithm proposed in this paper, termed as TV-$\mathcal{AB}$, thus is applicable to time-varying, directed graphs without the need of eigenvector estimation resulting from push-sum.

The TV-$\mathcal{AB}$ algorithm we introduce involves a unique and a rather counter-intuitive way of mixing information among the agents. As the graph at each iteration may not be strongly-connected, the mechanics of TV-$\mathcal{AB}$ can be explained over a single directed edge, $i \longrightarrow j$. First, we note that TV-$\mathcal{AB}$ involves two state updates: one with a row-stochastic weight matrix, $\mathcal{A}$, and the other with a column-stochastic weight matrix, $\mathcal{B}$. The update involving $\mathcal{A}$ is standard where the receiving agent j implements a sum-preserving update to its past and the incoming information from agent i , while agent i assigns a weight of 1 to its past since it does not receive any information. However, the additional update with the column-stochastic weights, $\mathcal{B}$, requires the transmitting agent i to implement a strictly stable update (by assigning a weight less than 1 to its past) and the receiving agent to implement an unstable update (by assigning weights that sum to > 1 to its past and the incoming information from agent i) in order to maintain column-stochasticity of $\mathcal{B}$. In other words, the updates involving $\mathcal{B}$ are not sum-preserving unlike the traditional information fusion.

We show that TV-$\mathcal{AB}$ converges linearly to the optimal solution when each local objective is smooth and the global objective is strongly-convex. The graph at each iteration can be generated randomly in an arbitrary fashion as long as the union of every C consecutive graphs is strongly-connected. This notion is known as bounded connectivity and is standard in the consensus and optimization literature on time-varying graphs, see e.g., [18], [26]. The bounded connectivity notion enables us to obtain more concrete convergence results as without this assumption, the analysis is restricted to the expected behavior of the optimization algorithm, see e.g., [39], [42], [44]. We show linear convergence with the help of a linear system of inequalities along with a matrix perturbation argument.

We now describe the rest of the paper. Section II formulates the problem and introduces the assumptions necessary to algorithm development. Section II-A develops and interprets the time-varying $\mathcal{AB}$ algorithm. Details on the convergence analysis are presented in Section III while Section IV states the main result. Finally, Section V provides numerical experiments and Section VI contains the concluding remarks.

Notation: We denote column vectors by lowercase bold letters, $\mathbf{x}$, and matrices by uppercase italics, X. The $n \times n$ identity matrix and the n-dimensional column vector of all ones are denoted by I_n and $\mathbf{1}_n$, respectively, where the dimension subscripts are dropped if clear from the context. We denote by $X \otimes Y$, the Kronecker product of two matrices, X and Y, while $\rho(X)$ denotes the spectral radius for a matrix, X. For a vector, $\mathbf{x}$, its i^{th} element is denoted by $[\mathbf{x}]_i$, while for a matrix, X, its $(i,j)^{\text{th}}$ element is denoted by $[X]_{i,j}$. The notation $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the spectral norm for matrices, whereas $\|\cdot\|_{\max}$ denotes the l_∞-norm on the set of square matrices.

II. Problem Formulation and Algorithm

In this section, we formulate the distributed optimization problem, state the assumptions, and introduce the TV-$\mathcal{AB}$ algorithm. To this aim, we assume that the set of agents, $\mathcal{V} = \{1, 2, \cdots, n\}$, communicate according to a time-varying directed graph, $\mathcal{G}_k(\mathcal{V}, \mathcal{E}_k)$, where k is the discrete-time index and $\mathcal{E}_k$ is the set of directed communication links at time k. An agent j can send information to an agent i, i.e., $j \longrightarrow i$, at time k, if and only if $(i,j) \in \mathcal{E}_k$. The goal of the agents is to collaboratively solve the following problem:

$$\text{Problem 1:} \qquad \min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{1}$$

where each local objective function, $f_i : \mathbb{R}^p \mapsto \mathbb{R}$, is held privately at agent i. We next formalize the set of assumptions that are standard in the distributed optimization literature.

Assumption 1 (Strong-convexity)**.** *The global objective function f, is μ-strongly-convex, i.e.,*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \tag{2}$$

for any $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^p$, where $\mu > 0$.

For Assumption 1 to hold it suffices that each f_i is convex and at least one of them is strongly-convex. Under this assumption, Problem 1 has a unique optimal solution, denoted by $\mathbf{x}^*$.

Assumption 2 (Smoothness)**.** *Each f_i is ℓ_i-smooth, i.e., it is differentiable and has a Lipschitz-continuous gradient. Mathematically, there exists $\ell_i > 0$ such that*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq \ell_i \|\mathbf{x} - \mathbf{y}\|_2, \tag{3}$$

for any $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^p$ and $\forall i \in \mathcal{V}$.

Assumption 2 implies that $f = \sum_i f_i$ is $\bar{\ell}$-smooth, where $\bar{\ell} = \frac{1}{n}\sum_{i=1}^n \ell_i$. Furthermore, collecting the local variables in column vectors, i.e.,

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^n \end{bmatrix}, \quad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}^1) \\ \vdots \\ f_n(\mathbf{x}^n) \end{bmatrix}, \quad \nabla\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \nabla f_1(\mathbf{x}^1) \\ \vdots \\ \nabla f_n(\mathbf{x}^n) \end{bmatrix},$$

we note that $\mathbf{f}$ is L-smooth, where $L = \max_i\{\ell_i\}$.

Assumption 3 (C-bounded strong-connectivity)**.** *For the sequence $\{\mathcal{G}_k = (\mathcal{V}, \mathcal{E}_k \subseteq \mathcal{V}\times\mathcal{V})\}$ of time-varying directed graphs, there exists some positive integer C such that the aggregate digraph $\mathcal{G}_k^C \triangleq (\mathcal{V}, \cup_{l=k}^{k+C-1}\mathcal{E}_l)$ is strongly-connected $\forall\, k \geq 0$.*

Assumption 4 (Weights)**.** *For the sequence $\{\mathcal{G}_k = (\mathcal{V}, \mathcal{E}_k)\}$ of time-varying directed graphs and the sequences, $\{A_k\}$ and $\{B_k\}$, of $n \times n$ matrices compliant with $\mathcal{G}_k$, i.e., $(i,j) \in \mathcal{E}_k \Leftrightarrow [A_k]_{i,j}, [B_k]_{i,j} \neq 0$, the following hold.*

(i) Stochasticity: $\{A_k\}$ and $\{B_k\}$ are row- and column-stochastic, respectively.

(ii) Aperiodicity: $\mathcal{G}_k$ has self-loops; i.e., $[A_k]_{i,i} > 0$ and $[B_k]_{i,i} > 0$, $\forall\, i \in \mathcal{V}$ and $\forall\, k \geq 0$.

(iii) Uniform positivity: There are scalars $0 < \alpha, \beta < 1$ such that $[A_k]_{i,j} \geq \alpha$ and $[B_k]_{i,j} \geq \beta$, $\forall\,(i,j) \in \mathcal{E}_k, k \geq 0$.

The strong-connectivity bound C introduced in Assumption 3 and the uniform positivity bounds α and β in Assumption 4 are not required to be known at any of the agents. They are only used in the analysis of the algorithm.

A. Algorithm Development

We now describe the TV-$\mathcal{AB}$ algorithm to solve Problem P1. At each time k, agent $i \in \mathcal{V}$ maintains two variables, $\mathbf{x}_k^i$, $\mathbf{y}_k^i$, both in $\mathbb{R}^p$, initialized with arbitrary $\mathbf{x}_0^i$ and $\mathbf{y}_0^i = \nabla f_i(\mathbf{x}_0^i)$. The $\mathbf{x}_k^i$-update at each agent is essentially gradient descent, albeit after mixing incoming information, and where the descent direction is given by an estimate of the global gradient, $\mathbf{y}_k^i$, instead of the local gradient, $\nabla f_i(\mathbf{x}_k^i)$. The $\mathbf{y}_k^i$-update at each agent tracks the global gradient and is based off of column-stochastic weights.

We first use a simple framework to explain the algorithm where only one edge $i \longrightarrow j$ is active at time k. The $\mathbf{x}_k$-update follows the standard sum-preserving notion where weights assigned to the past information are non-negative and sum to 1:

$$\begin{aligned} \mathbf{x}_{k+1}^i =& \mathbf{x}_k^i - \eta\mathbf{y}_k^i, \qquad \forall\, i \neq j, \\ \mathbf{x}_{k+1}^j =& [A_k]_{(j,i)}\mathbf{x}_k^i + [A_k]_{(j,j)}\mathbf{x}_k^j - \eta\mathbf{y}_k^j, \end{aligned}$$

where η is a constant step-size. The weight matrix, A_k, behind this update is row-stochastic: Each diagonal element, $[A_k]_{(i,i)} = 1, \forall\, i \neq j$, while the j^{th} row has only two positive elements such that $[A_k]_{(j,i)} + [A_k]_{(j,j)} = 1$.

Defining the auxiliary variable $\mathbf{z}_{k+1}^i$ as the successive difference, $\nabla f_i(\mathbf{x}_{k+1}^i) - \nabla f_i(\mathbf{x}_k^i)$, of the gradients, with $\mathbf{z}_0^i = \mathbf{0}_p$, the $\mathbf{y}_k$-update is given by

$$\begin{aligned} \mathbf{y}_{k+1}^m =& [B_k]_{(m,m)}\mathbf{y}_k^m + \mathbf{z}_{k+1}^m, \qquad m \neq i, m \neq j, \\ \mathbf{y}_{k+1}^i =& [B_k]_{(i,i)}\mathbf{y}_k^i + \mathbf{z}_{k+1}^i, \\ \mathbf{y}_{k+1}^j =& [B_k]_{(j,i)}\mathbf{y}_k^i + [B_k]_{(j,j)}\mathbf{y}_k^j + \mathbf{z}_{k+1}^j. \end{aligned}$$

Note that every column of B_k, except the i^{th}, has only one non-zero element as no agent other than i is transmitting. Hence, for B_k to be column-stochastic, $[B_k]_{(j,j)} = 1, \forall\, j \neq i$. For the transmitting agent, we have $[B_k]_{(i,i)} + [B_k]_{(j,i)} = 1$. In contrast to the traditional row- or doubly-stochastic updates, the $\mathbf{y}_{k+1}^i$-update is locally stable as $[B_k]_{(i,i)} < 1$, while the $\mathbf{y}_{k+1}^j$-update is locally unstable as $[B_k]_{(j,i)} + [B_k]_{(j,j)} > 1$.

B. The TV-$\mathcal{AB}$ Algorithm

The simple scenario discussed above can be generalized to arbitrary graphs, resulting into the following algorithm:

$$\mathbf{x}_{k+1}^i = \sum_{j=1}^n [A_k]_{i,j}\mathbf{x}_k^j - \eta\mathbf{y}_k^i, \tag{4a}$$

$$\mathbf{y}_{k+1}^i = \sum_{j=1}^n [B_k]_{i,j}\mathbf{y}_k^j + \mathbf{z}_k^i, \tag{4b}$$

where the weights $[A_k]_{i,j}$ and $[B_k]_{i,j}$ satisfy Assumption 4. Letting $\mathcal{A}_k \triangleq A_k \otimes I_p$ and $\mathcal{B}_k \triangleq B_k \otimes I_p$, we present Eqs. (4) in a vector-matrix form, where the local variables $\mathbf{x}_k^i$s and $\mathbf{y}_k^i$s and gradients $\nabla f_i(\mathbf{x}_k^i)$ are stored in the column vectors $\mathbf{x}_k$, $\mathbf{y}_k$, and $\nabla\mathbf{f}(\mathbf{x}_k)$, respectively:

$$\mathbf{x}_{k+1} = \mathcal{A}_k\mathbf{x}_k - \eta\mathbf{y}_k, \tag{5a}$$

$$\mathbf{y}_{k+1} = \mathcal{B}_k\mathbf{y}_k + \mathbf{z}_k, \tag{5b}$$

where $\mathbf{z}_k = \nabla\mathbf{f}(\mathbf{x}_k) - \nabla\mathbf{f}(\mathbf{x}_{k-1})$, and $\mathbf{z}_0 = \mathbf{0}_{np}$. The weight matrices, $\mathcal{A}_k$ and $\mathcal{B}_k$, are row- and column-stochastic, respectively. However, since each $\mathcal{G}_k$ is not necessarily strongly-connected, the weights $\mathcal{A}_k$ and $\mathcal{B}_k$ are not necessarily primitive or irreducible; thus, the standard Perron-Frobenius arguments are not applicable here. To overcome this isse, we use the notions of absolute probability sequences, ergodicity, and multi-step contractions, a recap of these concepts is provided in the Appendix: Section A.

III. Convergence Analysis

To proceed with the analysis, we perform a state transformation: $\mathbf{s}_k = (V_k^{-1} \otimes I_p)\mathbf{y}_k$ on $\mathbf{y}_k$, where $V_k = \text{diag}[\mathbf{v}_k]$ and $\mathbf{v}_k$ follows Eq. (6a). The TV-$\mathcal{AB}$ algorithm is thus equivalently written as

$$\mathbf{v}_{k+1} = B_k\mathbf{v}_k, \tag{6a}$$

$$\mathbf{x}_{k+1} = \mathcal{A}_k\mathbf{x}_k - \eta(V_k \otimes I_p)\mathbf{s}_k, \tag{6b}$$

$$\mathbf{s}_{k+1} = \mathcal{R}_k\mathbf{s}_k + (V_{k+1}^{-1} \otimes I_p)(\nabla\mathbf{f}(\mathbf{x}_{k+1}) - \nabla\mathbf{f}(\mathbf{x}_k)), \tag{6c}$$

where $\mathbf{v}_0 = \mathbf{1}_n$, $\mathcal{R}_k \triangleq R_k \otimes I_p$, and $R_k = V_{k+1}^{-1}B_kV_k$. It can be verified that $\{R_k\}$ is a sequence of row-stochastic matrices for which the absolute probability sequence is $\{\mathbf{v}_k\}$; see Appendix: Section A on absolute probability sequences.

We now proceed with the convergence analysis of the equivalent algorithm in Eqs. (6a)-(6c). Our approach rests on a few quantities that we describe next:

(i) $\overline{\mathbf{x}}_k^{\mathrm{w}} = (\boldsymbol{\phi}_k^\top \otimes I_p)\mathbf{x}_k$, which is the average of $\mathbf{x}_k^i$'s weighted by the absolute probability sequence, $\{\boldsymbol{\phi}_k\}$, of A_k's, see Corollary 1 in the Appendix: Section A;

(ii) $\widetilde{\mathbf{x}}_k^{\mathrm{w}} = \mathbf{x}_k - \mathbf{1}_n \otimes \overline{\mathbf{x}}_k^{\mathrm{w}}$, which can be regarded as the weighted consensus error in the network;

(iii) $\mathbf{r}_k = \mathbf{1}_n \otimes \overline{\mathbf{x}}_k^{\mathrm{w}} - \mathbf{1}_n \otimes \mathbf{x}^*$, which is the optimality gap associated with the weighted average;

(iv) $\widetilde{\mathbf{s}}_k^{\mathrm{w}} = \mathbf{s}_k - (\mathbf{1}_n \mathbf{v}_k^\top \otimes I_p)\mathbf{s}_k$, which is an error term corresponding to gradient estimation.

With the help of these quantities, we define the vector

$$\mathbf{t}_k = \begin{bmatrix} \|\widetilde{\mathbf{x}}_k^{\mathrm{w}}\|_2 \\ \|\mathbf{r}_k\|_2 \\ \|\widetilde{\mathbf{s}}_k^{\mathrm{w}}\|_2 \end{bmatrix}, \tag{7}$$

and show that it goes to zero as $k \to \infty$. Clearly, if $\mathbf{t}_k \to \mathbf{0}$, then $\mathbf{x}_k \to \mathbf{1}_n \otimes \mathbf{x}^*$, and rate of convergence of TV-$\mathcal{AB}$ is upper bounded by the rate at which $\mathbf{t}_k \to \mathbf{0}$. To establish that $\mathbf{t}_k \to \mathbf{0}$, we derive a linear system of inequalities that expresses the evolution of $\mathbf{t}_k$ in the following form:

$$\begin{bmatrix} \mathbf{t}_{k+1} \\ \vdots \\ \mathbf{t}_{k-(\overline{C}-2)} \end{bmatrix} \leq M(\eta) \begin{bmatrix} \mathbf{t}_k \\ \vdots \\ \mathbf{t}_{k-(\overline{C}-1)} \end{bmatrix}, \tag{8}$$

where the elements of $M(\eta)$ are the coefficients of the linear system. Clearly, if $\rho(M(\eta)) < 1$, then $\mathbf{x}_k \to \mathbf{1}_n \otimes \mathbf{x}^*$ at least at the rate of $\mathcal{O}(\rho(M(\eta))^k)$.

Fig. 1 provides a roadmap to establish Eq. (8). The next four lemmas provide the corresponding inequalities, whereas the proofs are deferred to the Appendix. The system of Eq. (8) is then analyzed in the next subsection.

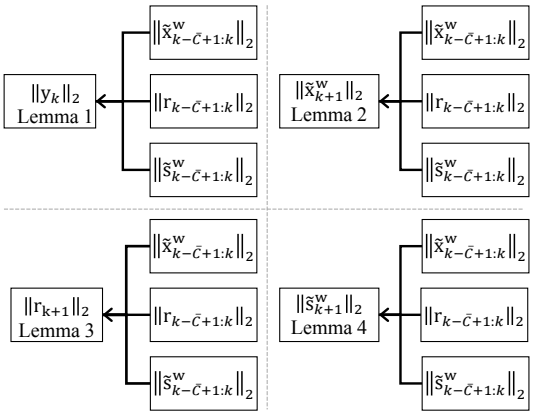

Fig. 1: Roadmap of deriving the linear system of inequalities.

Remark 1. *The constant $\overline{C} = \max\{\overline{C}_{\mathcal{A}}, \overline{C}_{\mathcal{B}}\}$ used in the rest of the paper, ensures concurrent multi-step contractions for both variables $\widetilde{\mathbf{x}}_k^w$ and $\widetilde{\mathbf{s}}_k^w$; see Lemma 7 and Corollary 2 in the Appendix: Section A for more details.*

Lemma 1. *The following inequality holds $\forall\, k \geq 0$:*

$$\|\mathbf{y}_k\|_2 \leq nL\|\widetilde{\mathbf{x}}_k^w\|_2 + nL\|\mathbf{r}_k\|_2 + \|\widetilde{\mathbf{s}}_k^w\|_2.$$

Proof: See Appendix: Section B. ■

Lemma 2. *The following inequality holds $\forall\, k \geq \overline{C} - 1$:*

$$\begin{aligned}\|\widetilde{\mathbf{x}}_{k+1}^w\|_2 \leq& \big(\gamma_{\mathcal{A}} + \eta Q_{\mathcal{A}} nL\big)\|\widetilde{\mathbf{x}}_{k-(\bar{C}-1)}^w\|_2 + \eta Q_{\mathcal{A}} nL \sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{x}}_{k-l}^w\|_2 \\ &+ \eta Q_{\mathcal{A}} nL\Big(\|\mathbf{r}_{k-(\bar{C}-1)}\|_2 + \sum_{l=0}^{\overline{C}-2} \|\mathbf{r}_{k-l}\|_2\Big) \\ &+ \eta Q_{\mathcal{A}}\Big(\|\widetilde{\mathbf{s}}_{k-(\bar{C}-1)}^w\|_2 + \sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{s}}_{k-l}^w\|_2\Big),\end{aligned}$$

where $\gamma_{\mathcal{A}}$ and $Q_{\mathcal{A}}$ are the constants defined in Lemma 7 in the Appendix: Section A.

Proof: See Appendix: Section C. ■

Lemma 3. *The following inequality holds $\forall\, k \geq 0$:*

$$\|\mathbf{r}_{k+1}\|_2 \leq \eta nL\|\widetilde{\mathbf{x}}_k^w\|_2 + \Big(1 - \eta\frac{\mu}{n^{nC-1}}\Big)\|\mathbf{r}_k\|_2 + \eta\sqrt{n}\|\widetilde{\mathbf{s}}_k^w\|_2.$$

Proof: See Appendix: Section D. ■

Lemma 4. *The following inequality holds $\forall\, k \geq \overline{C} - 1$:*

$$\begin{aligned}\|\widetilde{\mathbf{s}}_{k+1}^w\|_2 \leq& m\sqrt{n}(2+\eta L)\Big(\|\widetilde{\mathbf{x}}_{k-(\bar{C}-1)}^w\|_2 + \sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{x}}_{k-l}^w\|_2\Big) \\ &+ \eta mnL\Big(\|\mathbf{r}_{k-(\bar{C}-1)}\|_2 + \sum_{l=0}^{\overline{C}-2} \|\mathbf{r}_{k-l}\|_2\Big) \\ &+ (\eta m + \gamma_{\mathcal{B}})\|\widetilde{\mathbf{s}}_{k-(\bar{C}-1)}^w\|_2 + \eta m \sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{s}}_{k-l}^w\|_2,\end{aligned}$$

where $m = n^{nC} Q_{\mathcal{B}} L$, and $\gamma_{\mathcal{B}}$ and $Q_{\mathcal{B}}$ are the constants defined in Corollary 2 in the Appendix: Section A.

Proof: See Appendix: Section E. ■

A. The resulting linear system of inequalities

Summarizing the results of Lemmas 2-4, for $0 < \eta < \frac{2}{nL}$, Eq. (8) can be expanded as follows:

$$\begin{aligned}\mathbf{t}_{k+1} \leq& \underbrace{\begin{bmatrix} \eta Q_{\mathcal{A}} nL & \eta Q_{\mathcal{A}} nL & \eta Q_{\mathcal{A}} \\ \eta nL & 1 - \eta\frac{\mu}{n^{nC-1}} & \eta\sqrt{n} \\ m\sqrt{n}(2+\eta L) & \eta mnL & \eta m \end{bmatrix}}_{M_1} \mathbf{t}_k \\ &+ \underbrace{\begin{bmatrix} \eta Q_{\mathcal{A}} nL & \eta Q_{\mathcal{A}} nL & \eta Q_{\mathcal{A}} \\ 0 & 0 & 0 \\ m\sqrt{n}(2+\eta L) & \eta mnL & \eta m \end{bmatrix}}_{M_2} (\mathbf{t}_{k-1} + \ldots + \mathbf{t}_{k-(\overline{C}-2)}) \\ &+ \underbrace{\begin{bmatrix} \gamma_{\mathcal{A}} + \eta Q_{\mathcal{A}} nL & \eta Q_{\mathcal{A}} nL & \eta Q_{\mathcal{A}} \\ 0 & 0 & 0 \\ m\sqrt{n}(2+\eta L) & \eta mnL & \gamma_{\mathcal{B}} + \eta m \end{bmatrix}}_{M_{\overline{C}}} \mathbf{t}_{k-(\overline{C}-1)},\end{aligned}$$

which is equivalent to

$$\begin{bmatrix} \mathbf{t}_{k+1} \\ \mathbf{t}_k \\ \mathbf{t}_{k-1} \\ \vdots \\ \mathbf{t}_{k-(\overline{C}-2)} \end{bmatrix} \leq \begin{bmatrix} M_1 & M_2 & \cdots & M_2 & M_{\overline{C}} \\ I & & & & \\ & I & & & \\ & & \ddots & & \\ & & & I & \end{bmatrix} \begin{bmatrix} \mathbf{t}_k \\ \mathbf{t}_{k-1} \\ \vdots \\ \mathbf{t}_{k-(\overline{C}-2)} \\ \mathbf{t}_{k-(\overline{C}-1)} \end{bmatrix}. \tag{9}$$

The system matrix, $M(\eta)$, in the above can be partitioned as

$$\underbrace{\begin{bmatrix} M_1^0 & \cdots & M_2^0 & M_{\overline{C}}^0 \\ I & & & \\ & \ddots & & \\ & & I & \end{bmatrix}}_{M^0} + \eta \underbrace{\begin{bmatrix} M_1^E & \cdots & M_2^E & M_{\overline{C}}^E \\ \mathbf{0} & & & \\ & \ddots & & \\ & & \mathbf{0} & \end{bmatrix}}_{M^E}, \quad (10)$$

where

$$M_1^0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 2m\sqrt{n} & 0 & 0 \end{bmatrix}, \; M_1^E = \begin{bmatrix} Q_{\mathcal{A}} nL & Q_{\mathcal{A}} nL & Q_{\mathcal{A}} \\ nL & -\frac{\mu}{n^{nC-1}} & \sqrt{n} \\ m\sqrt{n}L & mnL & m \end{bmatrix},$$

$$M_2^0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2m\sqrt{n} & 0 & 0 \end{bmatrix}, \; M_2^E = \begin{bmatrix} Q_{\mathcal{A}} nL & Q_{\mathcal{A}} nL & Q_{\mathcal{A}} \\ 0 & 0 & 0 \\ m\sqrt{n}L & mnL & m \end{bmatrix},$$

$$M_{\overline{C}}^0 = \begin{bmatrix} \gamma_{\mathcal{A}} & 0 & 0 \\ 0 & 0 & 0 \\ 2m\sqrt{n} & 0 & \gamma_{\mathcal{B}} \end{bmatrix}, \; M_{\overline{C}}^E = \begin{bmatrix} Q_{\mathcal{A}} nL & Q_{\mathcal{A}} nL & Q_{\mathcal{A}} \\ 0 & 0 & 0 \\ m\sqrt{n}L & mnL & m \end{bmatrix}.$$

Recall that our goal is to establish the geometric decay of $\mathbf{t}_k$ in Eq. (7). To this purpose, it is sufficient to show $\rho(M(\eta)) < 1$. As a first step, we finish this section with a lemma on the spectral radius of the matrix M^0 in Eq. (10) and its corresponding eigenvector.

Lemma 5. *The spectral radius of the matrix M^0 is 1 and $\lambda = 1$ is a simple eigenvalue of M^0. The left and right eigenvectors, $M_0\mathbf{u} = \mathbf{u}, \mathbf{w}^\top M_0 = \mathbf{w}^\top$, are given by*

$$\mathbf{u} = \mathbf{1}_{\overline{C}} \otimes \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\top, \quad (11)$$

$$\mathbf{w}^\top = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}. \quad (12)$$

Proof: See Appendix: Section F. ∎

IV. Linear Convergence

We now state the main convergence result for TV-$\mathcal{AB}$.

Theorem 1. *The spectral radius of $M(\eta)$ is strictly less than 1 when η is sufficiently small. Therefore $\|\mathbf{x}_k - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$ converges to zero (at least) at the rate of $O\big(\rho(M(\eta))^k\big)$.*

Proof: From Lemma 5, let $q(\eta)$ be the simple eigenvalue of $M(\eta)$, as a function of η, for which $q(0) = 1$. Recall that $M(\eta)$ can be partitioned as $M^0 + \eta M^E$ from Eq. (10). Borrowing a result from matrix perturbation theory [52, Theorem 6.3.12], we have that

$$\left.\frac{dq(\eta)}{d\eta}\right|_{\eta=0} = \frac{\mathbf{w}^\top M^E \mathbf{u}}{\mathbf{w}^\top \mathbf{u}},$$

where $\mathbf{u}$ and $\mathbf{w}$ are right and left eigenvectors corresponding to the simple eigenvalue, $q(0)$. From Lemma 5, it can be verified that $\mathbf{w}^\top \mathbf{u} = 1$ and $\mathbf{w}^\top M^E \mathbf{u} = -\mu/n^{nC-1} < 0$, which implies that $\frac{d}{d\eta} q(\eta)$ is negative. Since the eigenvalues are a continuous function of the elements of a matrix, we have that $q(\eta)$ decreases for a sufficiently small η (slightly increasing from zero) and the theorem follows. ∎

V. Numerical Experiments

This section illustrates the application and performance of the time-varying $\mathcal{AB}$ algorithm in a variety of numerical experiments. In the rest of this section, we adopt a simple uniform weighting strategy to construct the row- and column-stochastic weights $[A_k]_{i,j}$ and $[B_k]_{i,j}$:

$$[A_k]_{i,j} = \begin{cases} 1/d^i_{k,\text{in}}, & (i,j) \in \mathcal{E}_k, \\ 0, & (i,j) \notin \mathcal{E}_k, \end{cases} \quad (13)$$

where $d^i_{k,\text{in}}$ is the in-degree of agent i at time k; and

$$[B_k]_{i,j} = \begin{cases} 1/d^j_{k,\text{out}}, & (i,j) \in \mathcal{E}_k, \\ 0, & (i,j) \notin \mathcal{E}_k, \end{cases} \quad (14)$$

where $d^j_{k,\text{out}}$ is the out-degree of agent j at time k.

A. Distributed binary classification

In this experiment, we study a binary classification problem using regularized logistic regression. Each agent i has access to m_i training samples: $(\mathbf{c}^{i(j)}, y^{i(j)}) \in \mathbb{R}^{p-1} \times \{-1, 1\}$, for $j = 1, 2, \cdots, m_i$, where $\mathbf{c}^{i(j)}$ is the $p-1$-dimensional feature vector of the j^{th} training sample at the i^{th} agent and $y^{i(j)}$ is the corresponding binary label. The agents collaboratively solve the following distributed logistic regression problem:

$$\min_{\mathbf{w},b} f(\mathbf{w}, b) = \sum_{i=1}^{n} f_i(\mathbf{w}, b),$$

where the private loss function f_i at agent i is

$$f_i(\mathbf{w}, b) = \sum_{j=1}^{m_i} \ln\left[1 + e^{\left(-\mathbf{w}^\top \mathbf{c}^{i(j)} + b\right) y^{i(j)}}\right] + \frac{\lambda}{2}\left(\|\mathbf{w}\|_2^2 + b^2\right).$$

The decision variable $\mathbf{w}$ represents the model weights assigned to the features and b is the bias term. It is straightforward to verify that the local loss functions f_i satisfy both Assumptions 1 and 2. The feature vectors, $\mathbf{c}^{i(j)}$, are drawn from IID Gaussian distributions with mean 0 and variance 9. We then generate binary labels from a Bernoulli distribution, with probability of $y^{i(j)} = +1$ being $(1 + e^{\widetilde{\mathbf{x}}^\top \mathbf{c}^{i(j)}})^{-1}$, where $\mathbf{w}$ and b are drawn from the standard uniform distribution. The network topology varies according to the periodic sequence of directed graphs as shown in Fig. 2 making the directed communication network 4-bounded strongly connected.

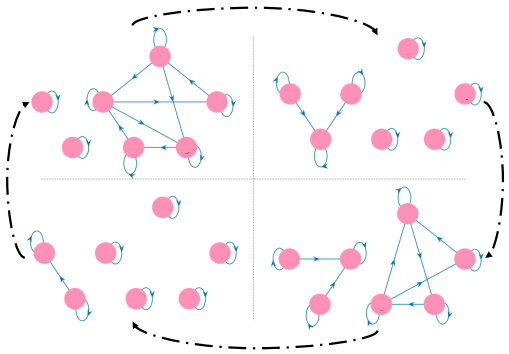

Fig. 2: Periodic time-varying topology, where the period is 4.

To solve the classification problem in a distributed manner, agents initialize their states $\mathbf{x}_0^i$ according to IID zero-mean Gaussian random variables with variance 9. The performance of TV-$\mathcal{AB}$ along with Push-DIGing [26], and subgradient-push [18] (with constant and diminishing step-sizes) is shown in Fig. 3 with the average residual $1/n\sum_{i=1}^n \|\mathbf{x}_k^i - \mathbf{x}^*\|_2$ as the evaluation metric. The step-sizes for TV-$\mathcal{AB}$ and Push-DIGing are hand-optimized. This numerical experiment confirms that time-varying $\mathcal{AB}$ converges linearly and is observed to be faster than Push-DIGing.

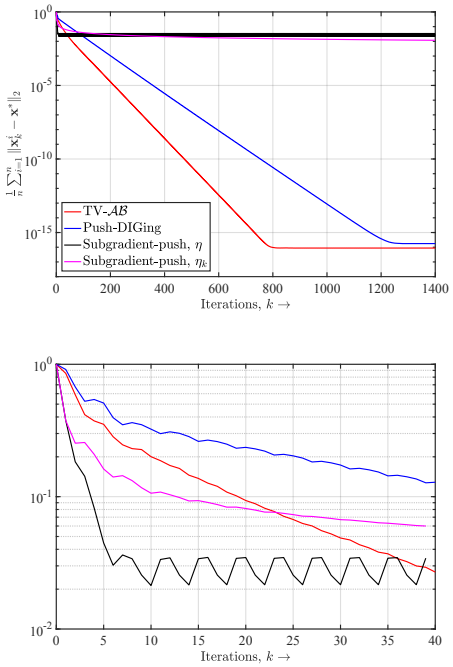

Fig. 3: Distributed logistic regression: Performance comparison with the transients magnified.

B. Distributed least-squares

In this example, $n = 60$ agents communicate according to a time-varying sequence of $C = 50$-bounded strongly-connected digraphs. The nodes are partitioned into 5 equally-sized clusters, and each cluster is internally strongly-connected at every iteration. The clusters communicate with each other according to a strongly-connected cluster-level network every $C = 50^{\text{th}}$ iteration, see Fig. 4. The agents aim to collaboratively find the solution $\mathbf{x}^*$ of the following least-squares problem:

$$f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^n \|H_i\mathbf{x} - \mathbf{b}_i\|_2^2,$$

where the vectors and matrices are of appropriate dimension. We choose each H_i such that it is rank-deficient but $\sum_i H_i^\top H_i$ is invertible. In other words, no agent can find $\mathbf{x}^*$ on its own and must cooperate.

To collaboratively solve the least-squares problem, agents initialize there states $\mathbf{x}_0^i$ according to IID standard Gaussian random variables. The performance of tTV-$\mathcal{AB}$ is compared with Push-DIGing and subgradient-push (with both constant and diminishing step-sizes) in Fig. 5 with the average residual $1/n\sum_{i=1}^n \|\mathbf{x}_k^i - \mathbf{x}^*\|_2$ as the comparison metric. The step-sizes follow a similar regime as in Experiment V-A. This numerical experiment, once again, confirms the linear convergence of $\mathcal{AB}$ to the optimal solution provided the step-size is sufficiently small.

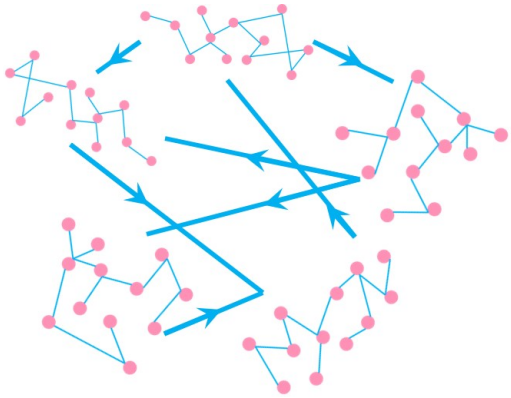

Fig. 4: $C = 50$-bounded strongly-connected digraph: The inter-cluster graph activates every 50th iteration.

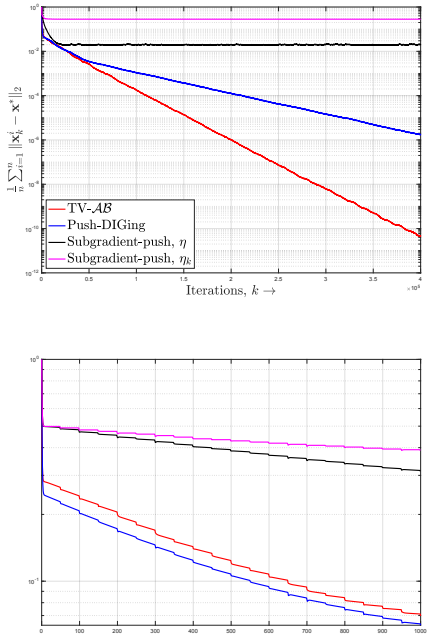

Fig. 5: Distributed least squares: Performance comparison with the transients magnified.

C. TV-$\mathcal{AB}$ on random graphs

We now apply TV-$\mathcal{AB}$ to random networks. In this scenario, $n = 80$ agents communicate over a $C = 15$-bounded strongly-connected network to solve the logistic regression problem of example V-A. The agents communicate over a strongly-connected random graph every 15th iteration and rely solely on local iterations for the rest of the time. The result of this experiment is presented in Fig. 6.

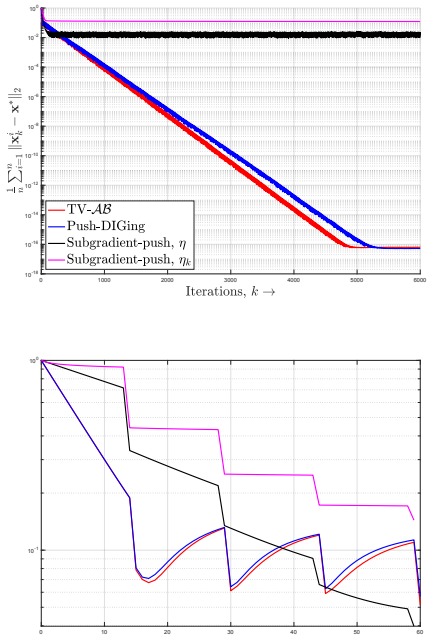

Fig. 6: Distributed logistic regression over random graphs: Performance comparison with the transients magnified.

In another scenario, 10 agents communicate according to the gossip protocol explained in Section II-A. The optimization problem is linear regression given 10 noisy samples of a line at each agent. The performance is compared in Fig. 7 with hand-optimized step-size.

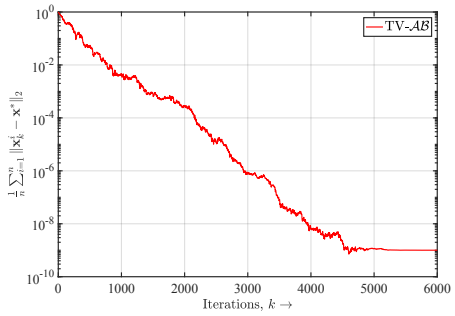

Fig. 7: Distributed linear regression : TV-$\mathcal{AB}$ under gossip.

VI. Conclusions

In this paper, we study TV-$\mathcal{AB}$ that minimizes a sum of smooth and strongly-convex functions over time-varying and possibly random directed graphs. We show that TV-$\mathcal{AB}$ converges linearly to the optimal solution when underlying time-varying graphs satisfy the standard bounded connectivity assumption, i.e., a union of every C consecutive graphs is strongly-connected. We derive the convergence result based on the stability analysis of a linear system of inequalities along with a matrix perturbation argument. We further provide extensive simulations that confirm the findings in this paper.

Appendix

A. Preliminaries

In this section, we recap some preliminaries that are used in the analysis. We start with the definitions of absolute probability sequences [53] and ergodicity [54].

Definition 1 (Absolute Probability Sequence)**.** *For row-stochastic matrices, $\{R_k\}$, an absolute probability sequence is a sequence $\{\boldsymbol{\pi}_k\}$ of stochastic vectors such that*

$$\boldsymbol{\pi}_k^\top = \boldsymbol{\pi}_{k+1}^\top R_k, \quad \forall k \geq 0.$$

Definition 2 (Ergodicity)**.** *A ergodic sequence of row-stochastic matrices, $\{R_k\}$, is such that for integers $p \geq 0$ and all $i, s = 1, \cdots, n$*

$$\lim_{c\to\infty} [U_{(c,p)}]_{i,s} \to d_s^p,$$

where $U_{(c,p)} = \Pi_{l=p}^{c} R_l$ is the backward product of $\{R_k\}$ and d_s^p is a constant not depending on i.

We next state a result on the ergodicity of the matrix sequence compliant with the aggregate digraph $\mathcal{G}_{sC}^C = (\mathcal{V}, \cup_{l=sC}^{sC+C-1} \mathcal{E}_l)$, see [55, Lemma 5.2.1] for details.

Lemma 6. *Under Assumptions 3 and 4, the row-stochastic matrix sequence $\{D_s = \Pi_{l=sC}^{sC+C-1} A_l\}$ compliant with the aggregate digraph $\mathcal{G}_{sC}^C = (\mathcal{V}, \cup_{l=sC}^{s+C-1} \mathcal{E}_l)$, for all $s \geq 0$, is ergodic, i.e.,*

$$\lim_{t\to\infty} D_t \cdots D_{s+1} D_s = \mathbf{1}\boldsymbol{\mu}_s^\top,$$

where $\{\boldsymbol{\mu}_s\}$ is the unique absolute probability sequence for $\{D_s\}$ (see, e.g., [53], [56, Lemma 1]) and is uniformly bounded away from zero, i.e., there exists $\delta \in (0, 1)$ such that $[\boldsymbol{\mu}_s]_i \geq \delta$, $\forall\, i \in \mathcal{V}$ and $\forall\, s \geq 0$. Furthermore, the convergence rate is geometric, i.e., $\forall\, t \geq s \geq 0$:

$$\|D_t \cdots D_{s+1} D_s - \mathbf{1}\boldsymbol{\mu}_s^\top\| \leq \mathcal{M} q^{t-s},$$

where the constants $\mathcal{M} > 0$ and $q \in (0, 1)$ depend only on n and α introduced in Assumption 4.

The next corollary extends the result above, deriving the absolute probability sequence for the sequence $\{A_k\}$ in terms of $\{\boldsymbol{\mu}_k\}$ in Lemma 6.

Corollary 1. *Under the assumptions of Lemma 6, the sequence $\{\boldsymbol{\phi}_k\}$ is an absolute probability sequence for the matrix sequence $\{A_k\}$, where*

$$\begin{aligned} \boldsymbol{\phi}_k^\top &= \boldsymbol{\mu}_k^\top, \quad k = sC, \\ \boldsymbol{\phi}_k^\top &= \boldsymbol{\mu}_{(s+1)C}^\top A_{(s+1)C-1} \cdots A_k, \quad k \in \big(sC, (s+1)C\big), \end{aligned} \tag{15}$$

for some $s \geq 0$.

Proof: Since the products $A_{(s+1)C-1} \cdots A_k$ are row-stochastic and $\{\boldsymbol{\mu}_s\}$ is an absolute probability sequence from Lemma 6, each $\boldsymbol{\mu}_s$ is a stochastic vector by definition and so is the vector $\boldsymbol{\phi}_k$ in Eq. (15). In fact,

$$\begin{aligned} \boldsymbol{\phi}_k^\top &= \boldsymbol{\mu}_{(s+1)C}^\top A_{(s+1)C-1} \cdots A_k, \\ &= \boldsymbol{\mu}_{(s+1)C}^\top A_{(s+1)C-1} \cdots A_{k+1}\ A_k = \boldsymbol{\phi}_{k+1}^\top A_k, \end{aligned}$$

i.e., the sequence $\{\boldsymbol{\phi}_k\}$ is an absolute probability sequence for $\{A_k\}$. ∎

The next lemma from [26, Lemma 5.3] establishes multi-step contraction of a backward product of a series of row-stochastic matrices $\{A_k\}$. This result is fundamental to the convergence analysis of TV-$\mathcal{AB}$.

Lemma 7 ($\overline{C}_{\mathcal{A}}$-step contraction for $\{\mathcal{A}_k\}$)**.** *Let Assumptions 3 and 4 hold. Recall that $\mathcal{A}_k = A_k \otimes I_p$ and define an integer $\overline{C}_{\mathcal{A}} \geq C$ such that*

$$\gamma_{\mathcal{A}} \triangleq Q_{\mathcal{A}}(1-\alpha^{nC})^{\frac{\overline{C}_{\mathcal{A}}-1}{nC}} < 1, \quad Q_{\mathcal{A}} = 2n\frac{1+\alpha^{-nC}}{1-\alpha^{nC}}. \tag{16}$$

Then for any $k \geq \overline{C}_{\mathcal{A}} - 1$ and any vector $\mathbf{b} \in \mathbb{R}^{np}$, if

$$\mathbf{a} = \mathcal{A}_{(\overline{C}_{\mathcal{A}},k-(\overline{C}_{\mathcal{A}}-1))}\mathbf{b},$$

where $\mathcal{A}_{(\overline{C}_{\mathcal{A}},k-(\overline{C}_{\mathcal{A}}-1))} \triangleq \mathcal{A}_k \cdots \mathcal{A}_{k-(\overline{C}_{\mathcal{A}}-1)}$, we have

$$\begin{aligned}&\big\|\big((I_n - \mathbf{1}_n\boldsymbol{\phi}_{k+1}^\top)\otimes I_p\big)\mathbf{a}\big\|_2\\ &\leq \gamma_{\mathcal{A}}\big\|\big((I_n - \mathbf{1}_n\boldsymbol{\phi}_{k-(\overline{C}_{\mathcal{A}}-1)}^\top)\otimes I_p\big)\mathbf{b}\big\|_2,\end{aligned}$$

where $\{\boldsymbol{\phi}_k\}$ is the absolute probability sequence of $\{A_k\}$, defined in Eq. (15)*.*

The next corollary establishes the multi-step contraction for the sequence $\{R_k\} = V_{k+1}^{-1}B_kV_k$, where $V_k = \text{diag}[\mathbf{v}_k]$ and $\{\mathbf{v}_k\}$ evolves as Eq. (6a).

Corollary 2 ($\overline{C}_{\mathcal{B}}$-step contraction for $\{\mathcal{R}_k\}$)**.** *Let Assumptions 3 and 4 hold. Recall that $\mathcal{R}_k = R_k \otimes I_p$ and define an integer $\overline{C}_{\mathcal{B}} \geq C$ such that*

$$\gamma_{\mathcal{B}} \triangleq Q_{\mathcal{B}}(1-\tau^{nC})^{\frac{\overline{C}_{\mathcal{B}}-1}{nC}} < 1, \quad Q_{\mathcal{B}} = 2n\frac{1+\tau^{-nC}}{1-\tau^{nC}}, \tag{17}$$

where $\tau = \frac{\beta}{n^{nC+1}}$; then for any $k \geq \overline{C}_{\mathcal{B}} - 1$ and any vector $\mathbf{b} \in \mathbb{R}^{np}$, if

$$\mathbf{a} = \mathcal{R}_{(\overline{C}_{\mathcal{B}},k-(\overline{C}_{\mathcal{B}}-1))}\mathbf{b},$$

where $\mathcal{R}_{(\overline{C}_{\mathcal{B}},k-(\overline{C}_{\mathcal{B}}-1))} = \mathcal{R}_k \cdots \mathcal{R}_{k-(\overline{C}_{\mathcal{B}}-1)}$, we have

$$\begin{aligned}&\big\|\big((I_n - \mathbf{1}_n\mathbf{v}_{k+1}^\top)\otimes I_p\big)\mathbf{a}\big\|_2\\ &\leq\gamma_{\mathcal{B}}\big\|\big((I_{np} - \mathbf{1}_n\mathbf{v}_{k-(\overline{C}_{\mathcal{B}}-1)}^\top)\otimes I_p\big)\mathbf{b}\big\|_2.\end{aligned}$$

Proof: It can be verified that

$$\begin{aligned}&((I_n - \mathbf{1}_n\mathbf{v}_{k+1}^\top)\otimes I_p)\mathcal{R}_{(\overline{C}_{\mathcal{B}},k-(\overline{C}_{\mathcal{B}}-1))}\\ =&((R_{(\overline{C}_{\mathcal{B}},k-(\overline{C}_{\mathcal{B}}-1))} - \mathbf{1}_n\mathbf{v}_{k-(\overline{C}_{\mathcal{B}}-1)}^\top)(I_n - \mathbf{1}_n\mathbf{v}_{k-(\overline{C}_{\mathcal{B}}-1)}^\top))\otimes I_p.\end{aligned}$$

Therefore,

$$\begin{aligned}&\big\|\big((I_n - \mathbf{1}_n\mathbf{v}_{k+1}^\top)\otimes I_p\big)\mathbf{a}\big\|_2\\ \leq&\big\|R_{(\overline{C}_{\mathcal{B}},k-(\overline{C}_{\mathcal{B}}-1))} - \mathbf{1}_n\mathbf{v}_{k-(\overline{C}_{\mathcal{B}}-1)}^\top\big\|_2\cdot\\ &\big\|((I_n - \mathbf{1}_n\mathbf{v}_{k-(\overline{C}_{\mathcal{B}}-1)}^\top)\otimes I_p)\mathbf{b}\big\|_2,\end{aligned}$$

where the inequality follows from the compatibility of vector 2-norm with matrix spectral norm. We now find an upper bound for the first term on the right hand side of the above equation. First note that $[R_k]_{i,j} = [B_k]_{ij}[\mathbf{v}_k]_j/[\mathbf{v}_{k+1}]_i$. From [18, Corollary 2(b)], we have that $[\mathbf{v}_k]_j \geq 1/n^{nC}$, $\forall\, k \geq 0$. Since $1/[\mathbf{v}_k]_j \geq 1/n$ and for any $(i,j) \in \mathcal{E}_k$, $[B_k]_{i,j} \geq \beta$ by Assumption 4, we have that in $[R_k]_{i,j} \geq \tau \triangleq \beta/n^{nC+1}$, for any $(i,j) \in \mathcal{E}_k$. Therefore, noting that for an $n \times n$ matrix, X, $\|X\|_2 \leq n\|X\|_{\max}$, we have

$$\begin{aligned}&\big\|R_{(\overline{C}_{\mathcal{B}},k-(\overline{C}_{\mathcal{B}}-1))} - \mathbf{1}_n\mathbf{v}_{k-(\overline{C}_{\mathcal{B}}-1)}^\top\big\|_2\\ \leq&n\big\|R_{(\overline{C}_{B},k-(\overline{C}_{B}-1))} - \mathbf{1}_n\mathbf{v}_{k-(\overline{C}_{B}-1)}^\top\big\|_{\max} \leq \gamma_{\mathcal{B}},\end{aligned}$$

where $\gamma_{\mathcal{B}} \triangleq 2n\frac{1+\tau^{-nC}}{1-\tau^{nC}}(1-\tau^{nC})^{\frac{\overline{C}_{\mathcal{B}}-1}{nC}}$ and the last inequality is from [11, Lemma 4(c)]. ∎

Finally, the next lemma is a standard result in optimization theory and states that the optimality gap in the domain space shrinks by at least a fixed ratio for a gradient descent step.

Lemma 8 ([57, Lemma 3.11])**.** *Let $g : \mathbb{R}^p \mapsto \mathbb{R}$ be μ-strongly-convex and have ℓ-Lipschitz gradient. Define $\mathbf{x}^+ = \mathbf{x} - \zeta\nabla g(\mathbf{x})$, where $0 < \zeta < 2/\ell$. Then*

$$\|\mathbf{x}^+ - \mathbf{x}^*\|_2 \leq \chi\|\mathbf{x} - \mathbf{x}^*\|_2$$

where $\chi = \max\{|1-\zeta\mu|, |1-\zeta\ell|\}$.

In the following, we provide the proofs of the lemmas described earlier in the paper.

B. Proof of Lemma 1.

Proof: Recall that $\mathbf{s}_k = \widetilde{\mathbf{s}}_k^{\mathrm{w}} + (\mathbf{1}_n\mathbf{v}_k^\top \otimes I_p)\mathbf{s}_k$ and it can be verified that

$$(\mathbf{1}_n\mathbf{v}_k^\top \otimes I_p)\mathbf{s}_k = (\mathbf{1}_n\mathbf{1}_n^\top \otimes I_p)\nabla\mathbf{f}(\mathbf{x}_k). \tag{18}$$

Exploiting the optimality condition $\nabla\mathbf{f}(\mathbf{1}_n \otimes \mathbf{x}^*) = \mathbf{0}_{np}$, we can express $\mathbf{s}_k$ as

$$\mathbf{s}_k = \widetilde{\mathbf{s}}_k^{\mathrm{w}} + (\mathbf{1}_n\mathbf{1}_n^\top \otimes I_p)(\nabla\mathbf{f}(\mathbf{x}_k) - \nabla\mathbf{f}(\mathbf{1}_n \otimes \mathbf{x}^*)).$$

Therefore, from the triangle inequality we have

$$\begin{aligned}\|\mathbf{s}_k\|_2 &\leq \|\widetilde{\mathbf{s}}_k^{\mathrm{w}}\|_2 + \sqrt{n}\textstyle\sum_{i=1}^n\|\nabla\mathbf{f}_i(\mathbf{x}_k^i) - \nabla\mathbf{f}_i(\mathbf{x}^*)\|_2,\\ &\leq \|\widetilde{\mathbf{s}}_k^{\mathrm{w}}\|_2 + \sqrt{n}L\textstyle\sum_{i=1}^n\|\mathbf{x}_k^i - \mathbf{x}^*\|_2,\\ &\leq \|\widetilde{\mathbf{s}}_k^{\mathrm{w}}\|_2 + nL\big\|\mathbf{x}_k + (\mathbf{1}_n\phi_k^\top \otimes I_p)(\mathbf{x}_k - \mathbf{x}_k) - \mathbf{1}_n \otimes \mathbf{x}^*\big\|_2,\\ &\leq \|\widetilde{\mathbf{s}}_k^{\mathrm{w}}\|_2 + nL\|\widetilde{\mathbf{x}}_k^{\mathrm{w}}\|_2 + nL\|\mathbf{r}_k\|_2.\end{aligned}$$

where the second inequality uses Lipschitz continuity and the third inequality is a consequence of Cauchy-Schwarz. Noting that $\mathbf{y}_k = (V_k \otimes I_p)\mathbf{s}_k$, we have $\|\mathbf{y}_k\|_2 \leq \|(V_k \otimes I_p)\|_2\|\mathbf{s}_k\|_2$ from the compatibility of vector 2-norm with matrix spectral norm. Next, since $\|(V_k \otimes I_p)\|_2 = \|V_k\|_2 = \max_i[\mathbf{v}_k]_i < 1$, we have

$$\|\mathbf{y}_k\|_2 \leq \|\mathbf{s}_k\|_2 \leq nL\|\widetilde{\mathbf{x}}_k^{\mathrm{w}}\|_2 + nL\|\mathbf{r}_k\|_2 + \|\widetilde{\mathbf{s}}_k^{\mathrm{w}}\|_2,$$

and the lemma follows. ∎

C. Proof of Lemma 2.

Proof: From Eq. (6b), we have

$$\mathbf{x}_{k+1} = \mathcal{A}_{(\overline{C},k-(\overline{C}-1))}\mathbf{x}_{k-(\overline{C}-1)} - \eta\sum_{l=0}^{\overline{C}-1}\mathcal{A}_{l,k-(l-1)}\mathbf{y}_{k-l},$$

which leads to

$$\begin{aligned}\|\widetilde{\mathbf{x}}_{k+1}^{\mathrm{w}}\|_2 =&\big\|\big((I_n - \mathbf{1}_n\boldsymbol{\phi}_{k+1}^\top)\otimes I_p\big)\mathbf{x}_{k+1}\big\|_2\\ \leq&\big\|\big((I_n - \mathbf{1}_n\boldsymbol{\phi}_{k+1}^\top)\otimes I_p\big)\mathcal{A}_{(\overline{C},k-(\overline{C}-1))}\mathbf{x}_{k-(\overline{C}-1)}\big\|_2\\ &+\eta\textstyle\sum_{l=0}^{\overline{C}-1}\big\|\big((I - \mathbf{1}_n\boldsymbol{\phi}_{k+1}^\top)\otimes I_p\big)\mathcal{A}_{l,k-(l-1)}\mathbf{y}_{k-l}\big\|_2.\end{aligned}$$

Consequently, $\forall\, k \geq \overline{C}-1$,

$$\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k+1}\|_2 \leq \gamma_{\mathcal{A}}\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-(\overline{C}-1)}\|_2 + \eta Q_{\mathcal{A}} \textstyle\sum_{l=0}^{\overline{C}-1} \|\mathbf{y}_{k-l}\|_2,$$

where the second term follows from [11, Lemma 4(c)]. From Lemma 1, we have

$$\begin{aligned}\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k+1}\|_2 \leq& (\gamma_{\mathcal{A}} + \eta Q_{\mathcal{A}} nL)\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-(\bar{C}-1)}\|_2,\\ &+ \eta Q_{\mathcal{A}} nL \textstyle\sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-l}\|_2,\\ &+ \eta Q_{\mathcal{A}} nL \|\mathbf{r}_{k-(\bar{C}-1)}\|_2 + \eta Q_{\mathcal{A}} nL \textstyle\sum_{l=0}^{\overline{C}-2} \|\mathbf{r}_{k-l}\|_2,\\ &+ \eta Q_{\mathcal{A}} \|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k-(\bar{C}-1)}\|_2 + \eta Q_{\mathcal{A}} \textstyle\sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k-l}\|_2,\end{aligned}$$

where $\gamma_{\mathcal{A}}$ and $Q_{\mathcal{A}}$ are the constants defined in Lemma 7. ■

D. Proof of Lemma 3.

Proof: Note that $\mathbf{y}_k - (\mathbf{v}_k \mathbf{1}_n^\top \otimes I_p)\mathbf{y}_k = (V_k \otimes I_p)\widetilde{\mathbf{s}}^{\mathrm{w}}_k$, we have

$$\begin{aligned}\|\mathbf{r}_{k+1}\|_2 &= \big\|(\mathbf{1}_n \boldsymbol{\phi}_{k+1}^\top \otimes I_p)\mathbf{x}_{k+1} - \mathbf{1}_n \otimes \mathbf{x}^*\big\|_2\\ &= \big\|(\mathbf{1}_n \boldsymbol{\phi}_{k+1}^\top \otimes I_p)\big(\mathcal{A}_k \mathbf{x}_k - \eta \mathbf{y}_k\\ &\quad + (\mathbf{v}_k \mathbf{1}_n^\top \otimes I_p)\mathbf{y}_k(-\eta + \eta)\big) - \mathbf{1}_n \otimes \mathbf{x}^*\big\|_2\\ &\leq \eta\big\|(\mathbf{1}_n \boldsymbol{\phi}_{k+1}^\top \otimes I_p)\big(\mathbf{y}_k - (\mathbf{v}_k \mathbf{1}_n^\top \otimes I_p)\mathbf{y}_k\big)\big\|_2\\ &\quad + \big\|(\mathbf{1}_n \boldsymbol{\phi}_k^\top \otimes I_p)\mathbf{x}_k - \mathbf{1}_n \otimes \mathbf{x}^* - \eta\theta_k(\mathbf{1}_n \mathbf{1}_n^\top \otimes I_p)\mathbf{y}_k\big\|_2\\ &\leq \eta\sqrt{n}\|\widetilde{\mathbf{s}}^{\mathrm{w}}_k\|_2 + \big\|(\mathbf{1}_n \boldsymbol{\phi}_k^\top \otimes I_p)\mathbf{x}_k - \mathbf{1}_n \otimes \mathbf{x}^*\\ &\qquad - \eta\theta_k(\mathbf{1}_n \mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(\mathbf{x}_k)\big\|_2, \qquad (19)\end{aligned}$$

where $\theta_k = \boldsymbol{\phi}_{k+1}^\top \mathbf{v}_k$. From [11, Lemma 4(c)], we note that $\theta_k \geq 1/n^{nC}$, while $\theta_k \leq 1, \forall k$, from Cauchy-Schwarz. The substitution of $(\mathbf{1}_n \mathbf{1}_n^\top \otimes I_p)\mathbf{y}_k$ by $(\mathbf{1}_n \mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(\mathbf{x}_k)$ follows a similar reasoning as in Eq. (18). Furthermore, the second term on the right hand side of the above inequality can be expressed as follows:

$$\begin{aligned}&\big\|\mathbf{1}_n \otimes \overline{\mathbf{x}}^{\mathrm{w}}_k - n\eta\theta_k \nabla \mathbf{f}(\mathbf{1}_n \otimes \overline{\mathbf{x}}^{\mathrm{w}}_k) - \mathbf{1}_n \otimes \mathbf{x}^{*\top}\\ &\quad + n\eta\theta_k \nabla \mathbf{f}(\mathbf{1}_n \otimes \overline{\mathbf{x}}^{\mathrm{w}}_k) - \eta\theta_k(\mathbf{1}_n \mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(\mathbf{x}_k)\big\|_2\\ \leq& \big\|\mathbf{1}_n \otimes \overline{\mathbf{x}}^{\mathrm{w}}_k - n\eta\theta_k \nabla \mathbf{f}(\mathbf{1}_n \otimes \overline{\mathbf{x}}^{\mathrm{w}}_k) - \mathbf{1}_n \otimes \mathbf{x}^{*\top}\big\|_2\\ &\quad + \eta\theta_k\big\|\mathbf{1}_n \otimes \big(\textstyle\sum_{i=1}^n (\nabla f_i(\overline{\mathbf{x}}^{\mathrm{w}}_k) - \nabla f_i(\mathbf{x}^i_k))\big)\big\|_2\\ \leq& \sqrt{n}\|\overline{\mathbf{x}}^{\mathrm{w}}_k - n\eta\theta_k \nabla f(\overline{\mathbf{x}}^{\mathrm{w}}_k) - \mathbf{x}^*\|_2 + \eta n L\|\overline{\mathbf{x}}^{\mathrm{w}}_k - \mathbf{x}_k\|_2,\end{aligned}$$

where the second term is due to Assumption 2. From Lemma 8, if $0 < \eta < \frac{2}{nL} < \frac{2}{nL\theta_k}$, we can bound the first term on the right hand side and obtain

$$\sqrt{n}\|\overline{\mathbf{x}}^{\mathrm{w}}_k - n\eta\theta_k \nabla f(\overline{\mathbf{x}}^{\mathrm{w}}_k) - \mathbf{x}^*\|_2 \leq \sqrt{n}\chi_k \|\overline{\mathbf{x}}^{\mathrm{w}}_k - \mathbf{x}^*\|_2 = \chi_k \|\mathbf{r}_k\|_2,$$

where

$$\begin{aligned}\chi_k =& \max\{|1 - n\eta\theta_k \mu|, |1 - n\eta\theta_k L|\}\\ =& 1 - n\eta\theta(k)\mu \leq 1 - \eta\frac{\mu}{n^{nC-1}}\end{aligned}$$

for $\eta < \frac{1}{nL} < \frac{1}{nL\theta_k}$. Going back to Eq. (19),

$$\begin{aligned}\|\mathbf{r}_{k+1}\|_2 \leq& \eta\sqrt{n}\|\widetilde{\mathbf{s}}^{\mathrm{w}}_k\|_2 + \chi_k \|\mathbf{r}_k\|_2 + \eta n L\|\widetilde{\mathbf{x}}^{\mathrm{w}}_k\|_2\\ \leq& \eta n L\|\widetilde{\mathbf{x}}^{\mathrm{w}}_k\|_2 + \big(1 - \eta\frac{\mu}{n^{nC-1}}\big)\|\mathbf{r}_k\|_2 + \eta\sqrt{n}\|\widetilde{\mathbf{s}}^{\mathrm{w}}_k\|_2,\end{aligned}$$

and the lemma follows. ■

E. Proof of Lemma 4.

Proof: From Eq. (6c), we have

$$\begin{aligned}\mathbf{s}_{k+1} =& \mathcal{R}_{(\overline{C},k-(\overline{C}-1))}\mathbf{s}_{k-(\overline{C}-1)}\\ &+ \sum_{l=0}^{\overline{C}-1} \mathcal{R}_{(l,k-(l-1))}(V^{-1}_{k-(l-1)} \otimes I_p)\mathbf{z}_{k-(l-1)}.\end{aligned}$$

Applying triangle inequality, we get

$$\begin{aligned}&\|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k+1}\|_2 = \|((I_n - \mathbf{1}_n \mathbf{v}_{k+1}^\top) \otimes I_p)\mathbf{s}_{k+1}\|_2\\ \leq& \|((I_n - \mathbf{1}_n \mathbf{v}_{k+1}^\top) \otimes I_p)\mathcal{R}_{(\overline{C},k-(\overline{C}-1))}\mathbf{s}_{k-(\overline{C}-1)}\|_2\\ &+ \sum_{l=-1}^{\overline{C}-2} \|((I_n - \mathbf{1}_n \mathbf{v}_{k+1}^\top) \otimes I_p)\mathcal{R}_{(l+1,k-l)}(V^{-1}_{k-l} \otimes I_p)\mathbf{z}_{k-l}\|_2,\end{aligned}$$

where $\mathcal{R}_{(l,k-l)} = I_{np}$ for $l \leq 0$. Therefore, $\forall\, k \geq \overline{C}-1$,

$$\begin{aligned}\|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k+1}\|_2 \leq& \gamma_{\mathcal{B}}\|((I_n - \mathbf{1}_n \mathbf{v}^\top_{k-(\overline{C}-1)}) \otimes I_p)\mathbf{s}_{k-(\overline{C}-1)}\|_2\\ &+ Q_{\mathcal{B}} \textstyle\sum_{l=-1}^{\overline{C}-2} \|(V^{-1}_{k-l} \otimes I_p)\mathbf{z}_{k-l}\|_2,\end{aligned}$$

where the second term follows [11, Lemma 4(c)]. With the help of [18, Corollary 2(b)] and Corollary 2, we have

$$\|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k+1}\|_2 \leq \gamma_{\mathcal{B}}\|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k-(\overline{C}-1)}\|_2 + n^{nC} Q_{\mathcal{B}} \textstyle\sum_{l=0}^{\overline{C}-1} \|\mathbf{z}_{k-(l-1)}\|_2.$$

From Assumption 2, the summation in the second term can be bounded as follows:

$$\textstyle\sum_{l=0}^{\overline{C}-1} \|\mathbf{z}_{k-(l-1)}\|_2 \leq L \sum_{l=0}^{\overline{C}-1} \|\mathbf{x}_{k-(l-1)} - \mathbf{x}_{k-l}\|_2.$$

Furthermore,

$$\begin{aligned}&\|\mathbf{x}_{k-(l-1)} - \mathbf{x}_{k-l}\|_2\\ =& \big\|(\mathcal{A}_{k-l} - I_{np})\big(\mathbf{x}_{k-l} - (\mathbf{1}_n \boldsymbol{\phi}_{k-l}^\top \otimes I_p)\mathbf{x}_{k-l}\big) - \eta\mathbf{y}_{k-l}\big\|_2\\ \leq& \|\mathcal{A}_{k-l} - I_{np}\|_2\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-l}\|_2 + \eta\|\mathbf{y}_{k-l}\|_2\end{aligned}$$

Summing over l leads to

$$\begin{aligned}\sum_{l=0}^{\overline{C}-1} \|\mathbf{x}_{k-(l-1)} - \mathbf{x}_{k-l}\|_2 \leq& 2\sqrt{n}\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-(\overline{C}-1)}\|_2 + \eta\|\mathbf{y}_{k-(\overline{C}-1)}\|_2\\ &+ \sum_{l=0}^{\overline{C}-2} \Big(2\sqrt{n}\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-l}\|_2 + \eta\|\mathbf{y}_{k-l}\|_2\Big).\end{aligned}$$

From Lemma 1, we have $\forall\, k \geq \overline{C}-1$:

$$\begin{aligned}\|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k+1}\|_2 \leq& m(2\sqrt{n} + \eta L n)\,(\|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-(\overline{C}-1)}\|_2 + \textstyle\sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{x}}^{\mathrm{w}}_{k-l}\|_2)\\ &+ \eta n m L\,(\|\mathbf{r}_{k-(\overline{C}-1)}\|_2 + \textstyle\sum_{l=0}^{\overline{C}-2} \|\mathbf{r}_{k-l}\|_2)\\ &+ (\eta m + \gamma_{\mathcal{B}})\|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k-(\overline{C}-1)}\|_2 + \eta m \textstyle\sum_{l=0}^{\overline{C}-2} \|\widetilde{\mathbf{s}}^{\mathrm{w}}_{k-l}\|_2,\end{aligned}$$

and the lemma follows. ■

F. Proof of Lemma 5

Proof: Based on the successive application of Schur's determinant identity [52], [58], the characteristic polynomial of M^0 is given by $(-1)^{\overline{C}}(\lambda - 1)(\gamma_{\mathcal{A}} - \lambda^{\overline{C}})(\gamma_{\mathcal{B}} - \lambda^{\overline{C}})\lambda^{(\overline{C}-1)}$. Given that $\gamma_{\mathcal{A}}, \gamma_{\mathcal{B}} \in (0, 1)$, the spectral radius of M^0 is 1 and the corresponding eigenvalue, $\lambda = 1$, is simple. We now proceed to find the corresponding right and left eigenvectors, $\mathbf{u}$ and $\mathbf{w}$, respectively.

By decomposing $\mathbf{u}$ and $\mathbf{w}$ as follows,

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}^1 \\ \vdots \\ \mathbf{u}^{\overline{C}} \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}^1 \\ \vdots \\ \mathbf{w}^{\overline{C}} \end{bmatrix},$$

where each $\mathbf{u}^i, \mathbf{w}^i$ is in $\mathbb{R}^3$, from $M^0\mathbf{u} = \mathbf{u}$, we get

$$\gamma_A[\mathbf{u}^{\overline{C}}]_1 = [\mathbf{u}^1]_1, \qquad \mathbf{u}^i = \mathbf{u}^{i+1}, \quad 1 \leq i \leq \overline{C} - 1,$$

resulting in $[\mathbf{u}^i]_1 = 0, \forall\, i$. Furthermore,

$$\gamma_B[\mathbf{u}^{\overline{C}}]_3 + 2m\sqrt{n}\sum_{i=1}^{\overline{C}}[\mathbf{u}^i]_1 = [\mathbf{u}^1]_3.$$

Therefore, $\gamma_B[\mathbf{u}^{\overline{C}}]_3 = [\mathbf{u}^1]_3$ which implies $[\mathbf{u}^i]_3 = 0, \forall\, i$. The entries $[\mathbf{u}^i]_2$ are free variables and we set them equal to $1, \forall\, i$. Consequently, $\mathbf{u}^i = \underline{\mathbf{u}} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\top$ and $\mathbf{u} = \mathbf{1}_{\overline{C}} \otimes \underline{\mathbf{u}}$.

Similarly, from $\mathbf{w}^\top M^0 = \mathbf{w}^\top$, we have

$$\begin{aligned} 2m\sqrt{n}[\mathbf{w}^1]_3 + [\mathbf{w}^{i+1}]_1 &= [\mathbf{w}^i]_1, \qquad i = 1, \ldots, \overline{C} - 1, \\ 2m\sqrt{n}[\mathbf{w}^1]_3 + \gamma_A[\mathbf{w}^1]_1 &= [\mathbf{w}^{\overline{C}}]_1. \end{aligned}$$

Summing over all i, we obtain

$$2\overline{C}m\sqrt{n}[\mathbf{w}^1]_3 + \gamma_A[\mathbf{w}^1]_1 = [\mathbf{w}^1]_1. \tag{20}$$

Furthermore,

$$\begin{aligned} [\mathbf{w}^1]_2 + [\mathbf{w}^2]_2 &= [\mathbf{w}^1]_2, \\ [\mathbf{w}^3]_2 &= [\mathbf{w}^2]_2, \\ &\vdots \\ [\mathbf{w}^{\overline{C}}]_2 &= [\mathbf{w}^{\overline{C}-1}]_2, \\ [\mathbf{w}^{\overline{C}}]_2 &= 0, \end{aligned}$$

resulting in $[\mathbf{w}^i]_2 = 0, \forall\, 2 \leq i \leq \overline{C}$. Note that $[\mathbf{w}^1]_2$ is a free variable and we can set it equal to 1. Additionally,

$$\begin{aligned} [\mathbf{w}^2]_3 &= [\mathbf{w}^1]_3, \\ [\mathbf{w}^3]_3 &= [\mathbf{w}^2]_3, \\ &\vdots \\ [\mathbf{w}^{\overline{C}}]_3 &= [\mathbf{w}^{\overline{C}-1}]_3, \\ \gamma_B[\mathbf{w}^1]_3 &= [\mathbf{w}^{\overline{C}}]_3, \end{aligned}$$

resulting in $[\mathbf{w}^i]_3 = 0$, which from Eq. (20) implies $[\mathbf{w}^i]_1 = 0$, for all i. Consequently,

$$\mathbf{w}^\top = \begin{bmatrix} 0 & [\mathbf{w}^1]_2 = 1 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

■

REFERENCES

[1] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[2] C. C. Moallemi and B. V. Roy, "Distributed optimization in adaptive networks," in *Proceedings of the 16th Advances in Neural Information Processing Systems Conference*, 2004, pp. 887–894.

[3] A. Agarwal, M. J. Wainwright, and J. C. Duchi, "Distributed dual averaging in networks," in *Proceedings of the 24th Advances in Neural Information Processing Systems Conference*, 2010, pp. 550–558.

[4] S. Shahrampour, S. Rakhlin, and A. Jadbabaie, "Online learning of dynamic parameters in social networks," in *Proceedings of the 27th Advances in Neural Information Processing Systems Conference*, 2013, pp. 2013–2021.

[5] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *3rd International Symposium on Information Processing in Sensor Networks*, Berkeley, CA, Apr. 2004, pp. 20–27.

[6] H. Raja and W. U. Bajwa, "Cloud K-SVD: A collaborative dictionary learning algorithm for big, distributed data," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 173–188, Jan. 2016.

[7] H.-T. Wai, Z. Yang, Z. Wang, and M. Hong, "Multi-agent reinforcement learning via double averaging primal-dual optimization," *arXiv preprint arXiv:1806.00877*, 2018.

[8] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning — Part I: Algorithm development," *arXiv preprint arXiv:1702.05122*, 2017.

[9] S. Safavi, U. A. Khan, S. Kar, and J. M. F. Moura, "Distributed localization: A linear theory," *Proceedings of the IEEE*, 2018.

[10] J. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[11] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[12] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic control*, vol. 57, no. 3, pp. 592–606, 2012.

[13] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, Sep. 2016.

[14] E. Wei and A. Ozdaglar, "Distributed alternating direction method of multipliers," in *51st IEEE Annual Conference on Decision and Control*, Dec. 2012, pp. 5445–5450.

[15] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Puschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2718–2723, May 2013.

[16] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *51st IEEE Annual Conference on Decision and Control*, Dec. 2012, pp. 5453–5458.

[17] K. I. Tsianos, "The role of the network in distributed optimization algorithms: Convergence rates, scalability, communication/computation tradeoffs and communication delays," Ph.D. dissertation, Dept. Elect. Comp. Eng. McGill University, 2013.

[18] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, Mar. 2015.

[19] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *44th Annual IEEE Symposium on Foundations of Computer Science*, Oct. 2003, pp. 482–491.

[20] F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *IEEE International Symposium on Information Theory*, Jun. 2010, pp. 1753–1757.

[21] C. Xi, Q. Wu, and U. A. Khan, "On the distributed optimization over directed networks," *Neurocomputing*, vol. 267, pp. 508–515, Dec. 2017.

[22] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3986–3992, Oct. 2016.

[23] K. Cai and H. Ishii, "Average consensus on general strongly connected digraphs," *Automatica*, vol. 48, no. 11, pp. 2750 – 2761, 2012.

[24] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *IEEE 54th Annual Conference on Decision and Control*, Dec. 2015, pp. 2055–2060.

[25] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, Apr. 2017.

[26] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, Dec. 2017.

[27] M. Zhu and S. Martínez, "Discrete-time dynamic average consensus," *Automatica*, vol. 46, no. 2, pp. 322–329, 2010.

[28] C. Xi, R. Xin, and U. A. Khan, "ADD-OPT: Accelerated distributed directed optimization," *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1329–1339, May 2018.

[29] C. Xi, V. S. Mai, R. Xin, E. Abed, and U. A. Khan, "Linear convergence in optimization over directed graphs with row-stochastic matrices," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3558–3565, Oct. 2018.

[30] R. Xin, C. Xi, and U. A. Khan, "FROST – Fast row-stochastic optimization with uncoordinated step-sizes," *Arxiv: https://arxiv.org/abs/1803.09169*, Mar. 2018.

[31] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[32] C. Xi and U. A. Khan, "DEXTRA: A fast algorithm for optimization over directed graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 4980–4993, Oct. 2017.

[33] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 325–330, Jul. 2018.

[34] G. Qu and N. Li, "Accelerated distributed Nesterov gradient descent," *Arxiv: https://arxiv.org/abs/1705.07176*, May 2017.

[35] P. R. N. Loizou, "Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descen methods," *Arxiv: https://arxiv.org/abs/1712.09677*, Dec. 2017.

[36] R. Xin and U. A. Khan, "Distributed heavy-ball: A generalization and acceleration of first-order methods with gradient tracking," *Arxiv: https://arxiv.org/abs/1808.02942*, Aug. 2018.

[37] Y. Tian, Y. Sun, B. Du, and G. Scutari, "ASY-SONATA: Achieving geometric convergence for distributed asynchronous optimization," *Arxiv: https://arxiv.org/pdf/1803.10359.pdf*, Mar. 2018.

[38] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, Apr. 2010.

[39] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1291–1306, Jun. 2011.

[40] A. Nedić and A. Olshevsky, "Distributed optimization of strongly convex functions on directed time-varying graphs," in *IEEE Global Conference on Signal and Information Processing*, Dec. 2013, pp. 329–332.

[41] Q. Lü, H. Li, and D. Xia, "Geometrical convergence rate for distributed optimization with time-varying directed graphs and uncoordinated step-sizes," *Information Sciences*, vol. 422, pp. 516–530, 2018.

[42] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of asynchronous distributed gradient methods over stochastic networks," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 434–448, 2018.

[43] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Mathematical Programming*, vol. 129, no. 2, pp. 255–284, 2011.

[44] D. Jakovetić, J. M. F. Xavier, and J. M. F. Moura, "Convergence rates of distributed Nesterov-like gradient methods on random networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 868–882, Feb. 2014.

[45] D. Jakovetić, D. Bajovic, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," *Arxiv: https://arxiv.org/abs/1803.07836*, Mar. 2018.

[46] M. Eisen, A. Mokhtari, and A. Ribeiro, "Decentralized quasi-Newton methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 10, pp. 2613–2628, May 2017.

[47] M. Powell, "Some global convergence properties of a variable metric algorithm for minimization without exact line searches," *Nonlinear Programming, SIAM-AMS Proceedings*, vol. 9, no. 1, 1976.

[48] R. Byrd, J. Nocedal, and Y. Yuan, "Global convergence of a class of quasi-Newton methods on convex problems," *SIAM Journal on Numerical Analysis*, vol. 24, no. 5, pp. 1171–1190, 1987.

[49] J. Nocedal and S. J. Wright, *Numerical optimization, 2^{nd} ed.* New York, NY: Springer series in operation research and financial engineering, 2006.

[50] T. Wu, K. Yuan, Q. Ling, W. Yin, and A. H. Sayed, "Decentralized consensus optimization with asynchrony and delays," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 2, pp. 293–307, June 2018.

[51] M. Assran and M. Rabbat, "Asynchronous subgradient-push," *arXiv preprint arXiv:1803.08950*, 2018.

[52] R. A. Horn and C. R. Johnson, *Matrix Analysis, 2^{nd} ed.* New York, NY: Cambridge University Press, 2013.

[53] A. Kolmogoroff, "Zur theorie der markoffschen ketten," *Mathematische Annalen*, vol. 112, pp. 155–160, 1936.

[54] E. Seneta, *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

[55] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D., Massachusetts Institute of Technology, Cambridge, MA, 1984.

[56] A. Nedić and J. Liu, "On convergence rate of weighted-averaging dynamics for consensus problems," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 766–781, Feb. 2017.

[57] S. Bubeck, "Convex optimization: Algorithms and complexity," *arXiv preprint arXiv:1405.4980*, 2014.

[58] P. D. Powell, "Calculating determinants of block matrices," *arXiv preprint arXiv:1112.4379*, 2011.